EXPERIMENTATION

Always-Valid Inference in Conversion Testing

A practitioner review of sequential testing methods and the statistical cost of peeking at A/B test results.

Bhaskar Roy Sarkar, cazyweb · June 28, 2026 ·

A/B testing · sequential testing · always-valid inference · peeking · conversion rate optimization

Download PDF

ABSTRACT

Conversion optimization teams routinely watch A/B tests in real time and stop them the moment a result looks significant. Under the fixed-horizon hypothesis tests that power most experimentation tooling, this practice, known as peeking, silently inflates the false-positive rate far above the nominal level, because the stopping decision is no longer independent of the data. This paper reviews the statistical machinery that lets practitioners monitor continuously without paying that penalty: classical sequential analysis and Wald's sequential probability ratio test, the group-sequential and alpha-spending boundaries developed for clinical trials, and the modern always-valid or anytime-valid framework built on mixture likelihood ratios, confidence sequences, and e-values. We trace how one commercial platform, Optimizely's Stats Engine, operationalized these ideas, and we close with concrete guidance for CRO teams choosing between fixed-horizon and sequential designs. Throughout, the emphasis is on what is

mathematically guaranteed versus what is merely convenient, so that practitioners can monitor honestly rather than peek dishonestly.

1. Introduction

Most ecommerce and CRO teams run A/B tests on platforms that report a p-value or a confidence interval, and most of them look at that number before the test is over. They watch a dashboard, see significance appear, and ship the winner. This behavior feels rational, but it breaks the assumption that the classical test rests on, and the result is a much higher rate of false wins than the reported significance level implies.

The problem is not new to statisticians. It is the same continuous-monitoring issue that motivated sequential analysis in the 1940s and group-sequential designs in clinical trials decades later, and it is the same flexibility-in-analysis issue that the methodological reform literature has flagged as a driver of irreproducible findings.[1] What is new is that online experimentation generates data continuously, decisions are made by non-statisticians, and the incentive to stop early is strong. This paper reviews the methods that let a team monitor as often as it likes while keeping its error guarantees intact: Wald's sequential probability ratio test (SPRT); group-sequential boundaries and alpha-spending; and the modern always-valid or anytime-valid framework of mixture SPRTs, confidence sequences, and e-values. We then describe how these ideas were deployed commercially, and we offer practical guidance.

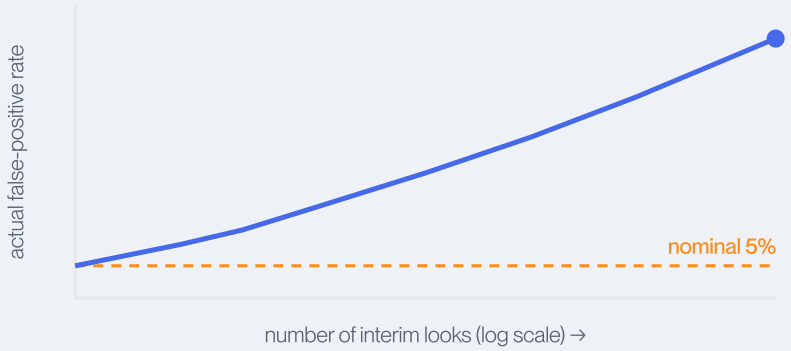
The American Statistical Association's statement on p-values is a useful anchor for the whole discussion: a p-value does not measure the probability that a hypothesis is true, and, critically, valid inference requires that the full data-collection and analysis process, including any stopping decision, be accounted for.[2] Peeking violates exactly that condition.

2. The Peeking Problem and Why It Inflates Type I Error

A fixed-horizon test is designed around a single, pre-committed sample size. You compute the sample size needed to detect a minimum effect at a chosen significance level (typically five percent) and power, you run until you reach it, and you test once. The five percent false-positive guarantee holds because there is exactly one opportunity to reject the null.

Peeking replaces that single test with many. If you evaluate significance after every visitor, or every day, and stop as soon as the p-value drops below your threshold, you are running a sequence of correlated tests and taking the most favorable one. Because the test statistic wanders randomly over time, even under a true null it will eventually cross any fixed threshold if you look often enough. The probability that it crosses at some point during the experiment is far larger than five percent. Johari and colleagues, working with Optimizely's customer data, formalized this for the A/B testing setting and showed both why the inflation occurs and how large it can be in practice, with continuous monitoring driving the true error rate well above the nominal level.[3] Their later journal treatment develops the same result with full rigor.[4]

Two clarifications matter for practitioners. First, the problem is a property of the decision rule, not of any single number being wrong; each individual p-value is computed correctly, but the rule that selects when to stop is what breaks the guarantee. Second, switching to a Bayesian posterior or a Bayes factor does not automatically rescue you. A Bayesian quantity can be valid under optional stopping, but only under specific stopping rules and priors; it is not a free pass, and naive continuous monitoring of a Bayesian credible interval can still mislead.[5] The standard practitioner reference on online experimentation makes the same warning a first-class concern.[6]



Schematic. With no real effect, repeatedly testing at a nominal 5% level and stopping at the first 'significant' look drives the true false-positive rate well above 5%. More looks mean more false winners.

3. Classical Sequential Analysis: Wald's SPRT

The foundational response to continuous data is Abraham Wald's sequential probability ratio test, introduced in his 1945 paper on sequential tests of statistical hypotheses.[7] Rather than fixing the sample size in advance, the SPRT accumulates the likelihood ratio between two simple hypotheses as each observation arrives, and it stops as soon as that ratio crosses one of two boundaries: an upper boundary to reject the null, a lower boundary to accept it, and a continuation region in between. The boundaries are chosen from the target Type I and Type II error rates.

The SPRT is not merely an early stopping heuristic; it is optimal in a precise sense. Wald and Wolfowitz proved that among all tests with the same error probabilities, the SPRT minimizes the expected number of observations under each of the two hypotheses.[8] For a CRO team, the practical reading is that sequential designs can reach a decision with fewer samples on average than a fixed-horizon test of equal error rates, which is why they are attractive when traffic is expensive or when shipping a winner early has real value.

The limitation is that the classical SPRT is built around two fully specified simple hypotheses, for example a known baseline rate versus a single known

alternative. Real conversion tests have a composite alternative: you do not know the true lift, only that you hope it is positive. Bridging from Wald's clean two-point setting to the composite, nonparametric reality of web experiments is the work that the rest of this review describes.



Wald's sequential probability ratio test. Accumulate the log-likelihood ratio of the data and stop the moment it crosses the upper (ship B) or lower (no effect) boundary; while it stays in between, keep collecting data.

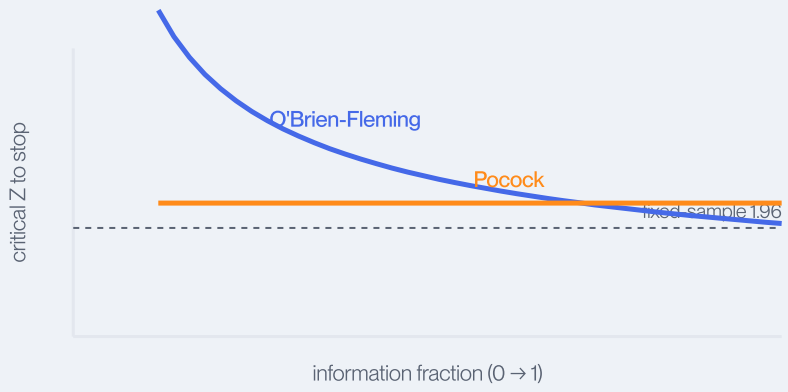
4. Group-Sequential Boundaries and Alpha-Spending

Clinical trials faced the peeking problem long before web experimentation existed, because ethics demand interim looks: you must be able to stop a trial early if one arm is clearly harming or clearly helping patients. The group-sequential literature solved this by allowing a pre-specified, finite number of interim analyses while keeping the overall Type I error at the target level.

Pocock proposed using a constant, more stringent significance threshold at every interim look, chosen so that the cumulative false-positive probability across all looks equals the target.[9] O'Brien and Fleming proposed an alternative shape in which the early looks use very conservative thresholds that relax as the trial proceeds, so that stopping early requires overwhelming evidence and the final analysis is tested at close to the nominal level.[10] These two boundary families embody a genuine tradeoff: Pocock spends error evenly

and stops earlier on average when effects are large, while O'Brien-Fleming protects the final analysis and is preferred when the planned full-sample test must retain near-full power.

The practical constraint in both schemes is that the number and timing of looks had to be fixed in advance. Lan and DeMets removed that rigidity with the alpha-spending function: instead of pre-committing to a schedule, you specify a function describing how the total error budget is spent as a function of the fraction of information accrued, and you may then analyze at flexible, even data-dependent, calendar times while still controlling the overall error.[11] The standard textbook treatment of these designs, including how to compute boundaries and expected sample sizes, is Jennison and Turnbull.[12] Group-sequential methods remain the right tool when you can tolerate a bounded, planned set of looks rather than literally continuous monitoring.



Group-sequential boundaries spend the 5% error budget across interim analyses. O'Brien-Fleming sets a very strict bar early and relaxes to near the fixed-sample cutoff (1.96) at the final look; Pocock holds a constant, higher bar throughout.

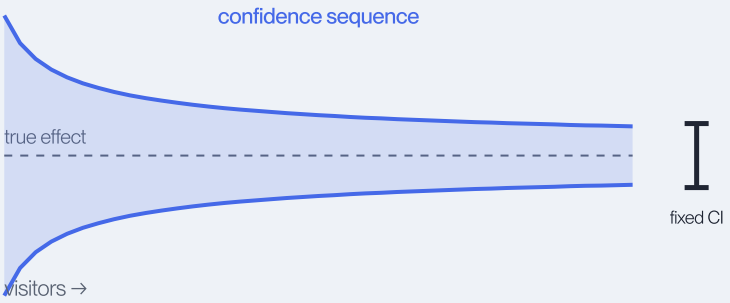
5. Always-Valid and Anytime-Valid Inference

The modern framework removes the last restriction: it permits inference that is valid at every sample size simultaneously, so you may look as often as you want, including after every single observation, and stop for any reason, without

inflating error. The core object is a p-value or confidence statement that is valid not just at one horizon but uniformly across time.

5.1 Confidence sequences

The estimation-side counterpart of a fixed-horizon confidence interval is the confidence sequence: a sequence of intervals that simultaneously cover the true parameter with the stated probability over the entire, unbounded run of the experiment. The concept dates to Robbins and collaborators, who connected it to the law of the iterated logarithm and constructed intervals valid for all sample sizes at once.[13] Because a confidence sequence holds at every time, you may stop the moment its interval excludes zero (or any null effect) and the coverage guarantee still holds. Howard, Ramdas, McAuliffe, and Sekhon substantially generalized this line of work, giving time-uniform, nonparametric, nonasymptotic confidence sequences whose widths shrink to zero and whose guarantees hold under broad conditions rather than only for clean parametric models, which is what makes them usable on real conversion and revenue metrics.[14]



A confidence sequence (left) holds its coverage at every sample size at once, so you may stop and look whenever you like and still trust the interval. A fixed-horizon interval (right) is only valid at the single pre-planned endpoint.

5.2 Mixture SPRT

The testing-side bridge from Wald's simple-versus-simple SPRT to the composite alternatives of real experiments is the mixture sequential probability ratio test (mSPRT). Rather than committing to one alternative value of the lift, the mSPRT integrates the likelihood ratio against a mixing distribution (a prior) over possible effect sizes, producing a single test martingale that yields an always-valid p-value at every observation. This is the engine behind the always-valid inference framework of Johari and colleagues, who use the mSPRT to construct always-valid p-values and confidence intervals for continuously monitored A/B tests and to extend the guarantee to multiple-comparison settings.[4]

5.3 E-values and the betting interpretation

A newer and unifying perspective casts the whole enterprise in terms of e-values and e-processes. An e-value is a nonnegative statistic whose expected value under the null is at most one; an e-process is a sequence of such statistics that respects optional stopping. The intuition is a fair bet against the null: if you bet repeatedly and your wealth grows large, that growth is itself evidence against the null, and because no fair betting strategy can be expected to make money against a true null, the evidence is valid no matter when you choose to cash out.[15] E-values multiply cleanly across independent studies and remain valid under optional continuation, which is exactly the property that p-values lack.[16] Ramdas, Grunwald, Vovk, and Shafer give a comprehensive synthesis of this game-theoretic approach and show that safe anytime-valid inference, e-processes for testing and confidence sequences for estimation, provides a single coherent foundation for monitoring accumulating data and stopping at any time.[17]

These are not only theoretical advances. Industry experimentation teams have built anytime-valid methods into production. Lindon and colleagues at Netflix developed anytime-valid confidence sequences and sequential tests for the linear model and for regression-adjusted treatment effects, explicitly to let software experiments be monitored continuously and stopped early while preserving coverage.[18]

6. How Commercial Tools Implement This

The clearest commercial case is Optimizely's Stats Engine, which moved the platform off fixed-horizon testing and onto sequential, always-valid inference. The publicly described design uses the mixture SPRT to produce always-valid p-values so that customers, who were already peeking, could be given results that stay valid at every look, and it pairs this with false-discovery-rate control across the many goals and variations a user tests at once.[19] The peer-reviewed account of the underlying methodology is the always-valid inference work, which reports both the statistical construction and its deployment.[4] The same intellectual lineage, mSPRT and confidence sequences for testing, plus e-processes as the unifying language, now appears across multiple large experimentation platforms.[17]

It is worth being precise about what these tools guarantee. An always-valid p-value controls the probability of a false rejection at any stopping time; it does not promise that a sequential test reaches significance as fast as a correctly sized fixed-horizon test would at its single planned endpoint. Sequential validity is bought with a modest efficiency cost when the true effect happens to match the fixed design's assumptions, and repaid when the effect is larger than planned (you stop early) or when you would otherwise have peeked (you avoid a false win). The honest framing for a practitioner is that always-valid methods price in the monitoring you were going to do anyway.

7. Practical Guidance for CRO Teams

First, decide your monitoring discipline before the test, not during it. If your tooling reports a fixed-horizon p-value, the only defensible practice is to compute a sample size in advance and test once at the planned horizon; intermediate looks for curiosity are fine, but the decision must wait.[6] If you cannot resist acting on interim results, you should be using a method designed for it.

Second, match the method to how you actually work. If you can commit to a small, planned number of interim analyses, a group-sequential design with an alpha-spending function gives you early-stopping ability with well-understood power, and O'Brien-Fleming-style boundaries are a reasonable default when you want to protect the final analysis.[1110] If you genuinely want to watch continuously and stop whenever the data justify it, use an always-valid or anytime-valid method: a confidence sequence for estimating the lift, or an mSPRT-based always-valid p-value for the decision.[144]

Third, do not treat Bayesian dashboards as a loophole. Continuous monitoring of a posterior is only safe under specific conditions, and the burden is on you to verify them rather than assume them.[5]

Fourth, remember the broader inference hygiene that no sequential method fixes: a valid p-value is still not the probability that your variant wins, multiple goals and segments still require multiplicity control, and the cleanest experiment can still be undone by instrumentation bugs or sample-ratio mismatch.[26] Always-valid inference solves the peeking problem specifically; it does not absolve you of trustworthy experiment design.

8. Conclusion

Peeking is not a moral failing of impatient marketers; it is a predictable response to data that arrives continuously, and under fixed-horizon statistics it quietly inflates false positives. The discipline of sequential analysis, from Wald's SPRT through group-sequential and alpha-spending designs to today's confidence sequences and e-values, exists precisely to make continuous monitoring honest. For CRO teams, the takeaway is simple to state and worth taking seriously: if you are going to look, use a method that is valid every time you look, and be clear-eyed about what that validity does and does not buy you.

References

1. Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*. https://doi.org/10.1177/0956797611417632

2. Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*. https://doi.org/10.1080/00031305.2016.1154108

3. Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2017). Peeking at A/B Tests: Why It Matters, and What to Do About It. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. https://doi.org/10.1145/3097983.3097992

4. Johari, R., Pekelis, L., Koomen, P., & Walsh, D. J. (2022). Always Valid Inference: Continuous Monitoring of A/B Tests. *Operations Research*. https://doi.org/10.1287/opre.2021.2135

5. Deng, A., Lu, J., & Chen, S. (2016). Continuous Monitoring of A/B Tests without Pain: Optional Stopping in Bayesian Testing. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. https://doi.org/10.1109/DSAA.2016.33

6. Kohavi, R., Tang, D., & Xu, Y. (2020). Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. *Cambridge University Press*. https://doi.org/10.1017/9781108653985

7. Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*. https://doi.org/10.1214/aoms/1177731118

8. Wald, A., & Wolfowitz, J. (1948). Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics*. https://doi.org/10.1214/aoms/1177730197

9. Pocock, S. J. (1977). Group Sequential Methods in the Design and Analysis of Clinical Trials. *Biometrika*. https://doi.org/10.1093/biomet/64.2.191

10. O'Brien, P. C., & Fleming, T. R. (1979). A Multiple Testing Procedure for Clinical Trials. *Biometrics*. https://doi.org/10.2307/2530245

11. Lan, K. K. G., & DeMets, D. L. (1983). Discrete Sequential Boundaries for Clinical Trials. *Biometrika*. https://doi.org/10.1093/biomet/70.3.659

12. Jennison, C., & Turnbull, B. W. (2000). Group Sequential Methods with Applications to Clinical Trials. *Chapman & Hall/CRC*. https://www.routledge.com/Group-Sequential-Methods-with-Applications-to-Clinical-Trials/Jennison-Turnbull/p/book/9780849303166

13. Robbins, H. (1970). Statistical Methods Related to the Law of the Iterated Logarithm. *The Annals of Mathematical Statistics*. https://doi.org/10.1214/aoms/1177696786

14. Howard, S. R., Ramdas, A., McAuliffe, J., & Sekhon, J. (2021). Time-Uniform, Nonparametric, Nonasymptotic Confidence Sequences. *The Annals of Statistics*. https://doi.org/10.1214/20-AOS1991

15. Shafer, G. (2021). Testing by Betting: A Strategy for Statistical and Scientific Communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. https://doi.org/10.1111/rssa.12647

16. Grunwald, P. D., de Heide, R., & Koolen, W. M. (2024). Safe Testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. https://doi.org/10.1093/jrsssb/qkae011

17. Ramdas, A., Grunwald, P., Vovk, V., & Shafer, G. (2023). Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*. https://doi.org/10.1214/23-STS894

18. Lindon, M., Ham, D. W., Tingley, M., & Bojinov, I. (2024). Anytime-Valid Inference in Linear Models and Regression-Adjusted Causal Inference. *Harvard Business School Working Paper No. 24-060*. https://arxiv.org/abs/2210.08589

19. Pekelis, L. (2015). The Story Behind Our Stats Engine. *Optimizely Blog*. https://www.optimizely.com/insights/blog/statistics-for-the-internet-age-the-story-behind-optimizelys-new-stats-engine/

Cite this

APA Copy

```
Bhaskar Roy Sarkar (2026). Always-Valid Inference in Conversion
Testing. cazyweb Research. https://cazyweb.com/research/always-valid-
inference-conversion-testing
```

BIBTEX Copy

```
@techreport{always-valid-inference-conversion-testing,
  author = {Bhaskar Roy Sarkar},
  title = {Always-Valid Inference in Conversion Testing},
  institution = {cazyweb Research},
  year = {2026},
  url = {https://cazyweb.com/research/always-valid-inference-
conversion-testing}
}
```

RELATED READING

Statistical significance in A/B testing, explained

How long should you run an A/B test?

Want a team to run experiments like this for you? See how we help

Let's build something that converts.

Start a project with us, or join the bench of specialists and agencies we deliver with.

Start a project ↗

cazyweb

The engineering and growth team behind leading brands and companies across the US and Europe.

CRO and ecommerce insights

Occasional, genuinely useful notes. No fluff.

you@company.com Subscribe →

Work email only. No spam. Unsubscribe any time. Privacy.

Solutions

Product Optimization

For Ecommerce

CRO Program

Experimentation Program

cazyweb Platform

Email & Retention

Paid Ads

Landing Pages

Revenue Calculator FREE

A/B Test Calculator FREE

CRO Agency

Shopify CRO

Ecommerce Growth

Website Performance

Company

About

For Brands

For Agencies

Careers HIRING

Contact

Explore

Services

Pricing

Performance program

Subscription

Website + CRO from $99

Case Studies

Blog

CRO Ideas

Apps

Learn

CazyWeb IT Solutions

GSTIN: 19JTLPS8476P2ZZ

hello@cazyweb.com

© 2026 cazyweb. All rights reserved.